Drug-Target Interaction Networks Prediction using Short-linear Motifs

Wenxiao Xu, Luis Rueda, Alioune Ngom

School of Computer Science, University of Windsor, Windsor, Ontario, Canada

Abstract

Drug-target interaction (DTI) prediction is a fundamental step in drug discovery and treatment of disease. Given a drug-target pair (d_i, t_j) with a drug compound d_i and a target protein t_j, our task is to assign a positive class label 1 to pair (d_i, t_j) if d_i interacts with t_j, and a negative class label 0 if they do not interact. We use short linear motifs (SLiMs) [1] as protein features and chemical substructure fingerprints [2] as drug features, and combine these features into a feature vector representing a drug-target pair. Given a DTI network, we represent all interacting drug-target pairs as feature vectors and use them as our positive class, and then devise a method to construct a negative class (i.e., a set of non-interacting pairs). Given a data set constructed as discussed above, we apply the feature selection method mRMR technique and the classification methods Support Vector Machine and Random Forest with ten-fold cross-validation to predict DTIs. Our preliminary results on four benchmark data sets yields higher AUC values when compared to current state-of-the-art DTI prediction methods.

Introduction

Known DTI networks can be obtained from KEGG BRITE, DrugBank, BRENDA. However, current DTI networks are relatively small, and contain too many unlabeled drug-target pairs; i.e., the absence of edge in a pair (d_i, t_j) means an unknown class, rather than a known true negative class. Given a arbitrary pair, (d_i, t_j), the main goal is to predict whether d_i interacts with t_j or not. Machine learning (ML) methods are more efficient than biochemical experiments given the existence of many DTI network databases. ML-based methods can be divided into two types: similarity-based and feature-based methods.

Similarity-based approach: uses drug-drug and protein-protein similarity matrices to predict DTIs.

Feature-based approach: uses descriptors of both proteins and drugs to predict DTIs.

Additionally, only positive data can be obtained from DTI networks. Existing methods consider the absence of interaction between a pair of drug and protein in DTI networks as a true negative interaction. However, this is incorrect. Hence, we propose a new feature-based approach and devise a strategy of constructing a negative data.

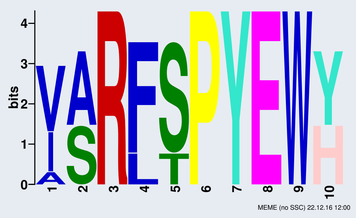

Figure 1: A motif in protein sequence

Materials and Methods

Gold Standard Data

Enzyme, Ion Channel, GPCR, Nuclear Receptor. [3]

Methods

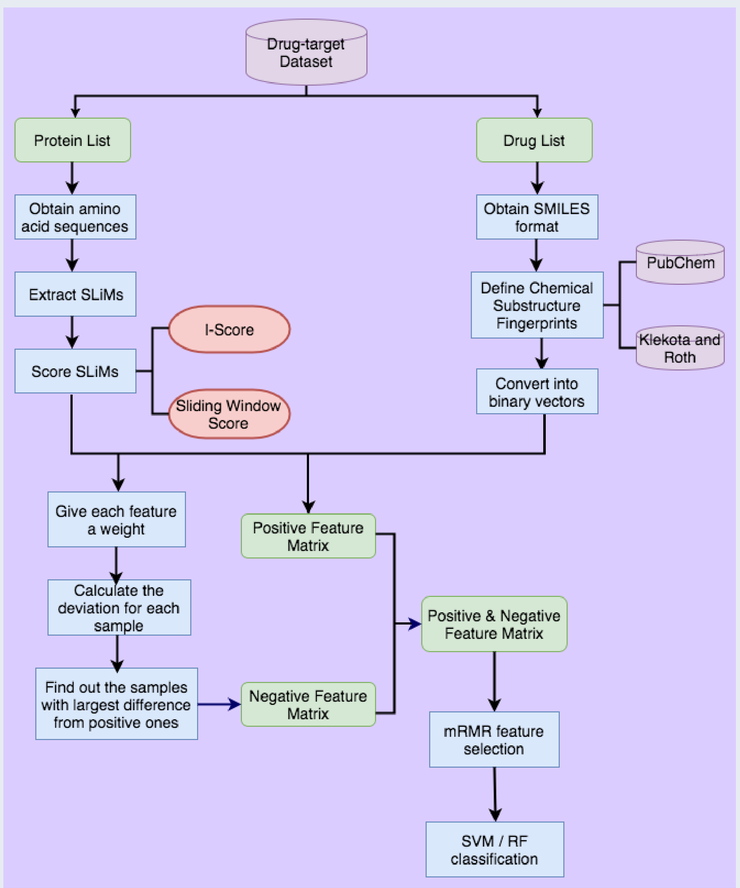

Figure 2: Methodology Flow Chart

Resuts and Comparison

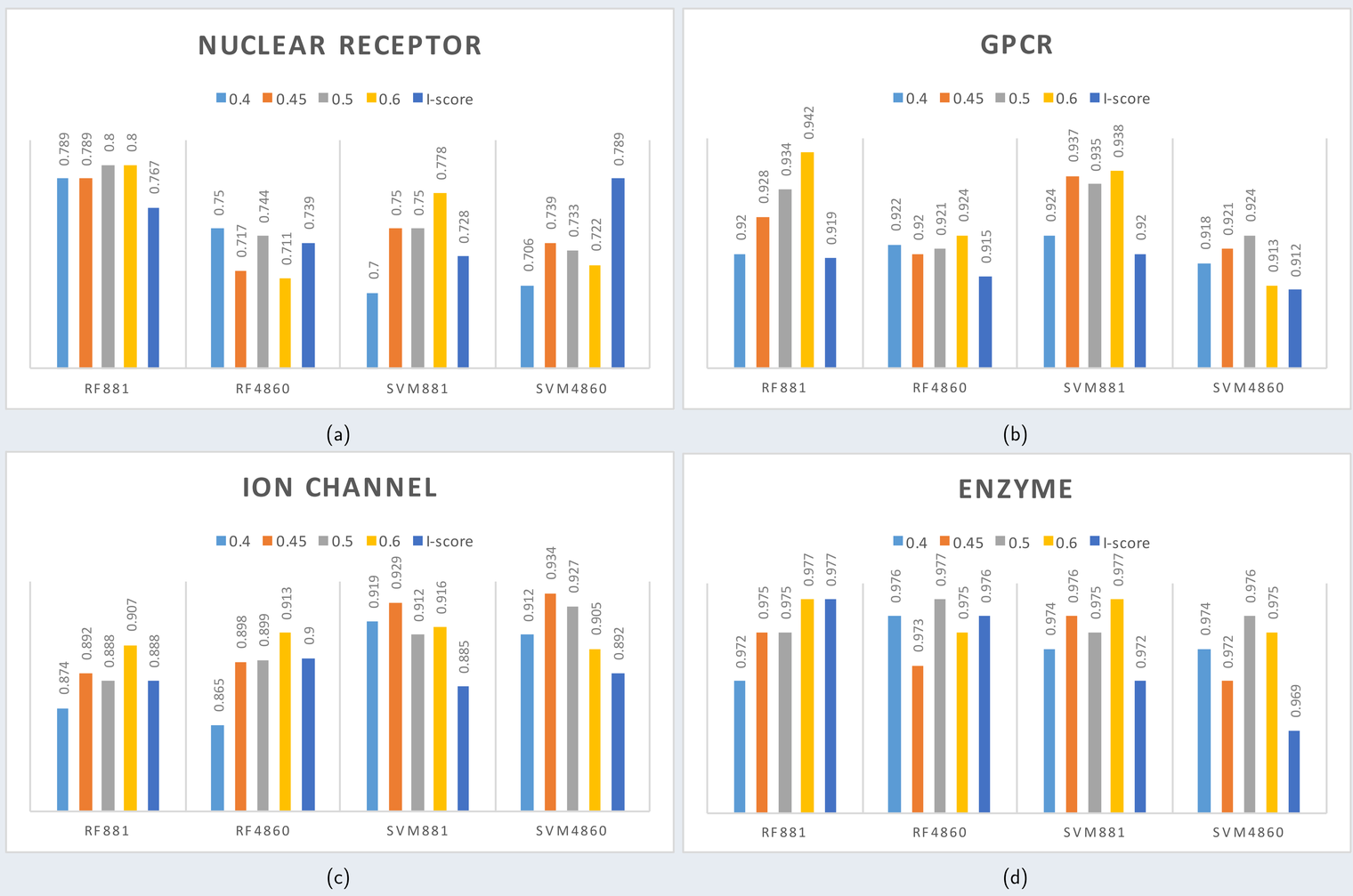

Figure 3: Accuracy values of four thresholds (0.4, 0.45, 0.5, 06) under SWS method, different SLiMs scoring methods (SWS and I-score), two classifiers (RF and SVM), different types of fingerprints defined by PubChem (881) and Klekota and Roth (4860) for each dataset.

Table 1 lists the AUC values of some existing methods using the same gold standard dataset, which are Cao et al. (2012) [4], Bigram-PSSM [5], Yamanishi et al. (2008) [3], Wang et al. (2010) [6], Yamanishi et al. (2010) [7], KBMF2K [8], NetCBP [9], and DBSI [10].

Table 1: The comparison of AUC among existing methods using benchmark datasets

Algorithms	Enzyme	Ion Channel	GPCR	Nuclear Receptor
Proposed Method	**0.9904**	**0.9639**	**0.9733**	0.8764
Cao et al. (2012)	0.9486	0.9428	0.8902	**0.8822**
Bigram-PSSM	0.948	0.889	0.872	0.869
Yamanishi et al. (2008)	0.904	0.851	0.899	0.835
Wang et al. (2010)	0.886	0.893	0.873	0.824
Yamanishi et al. (2010)	0.892	0.812	0.827	0.835
KBMF2K	0.832	0.799	0.857	0.824
NetCBP	0.8251	0.8034	0.8235	0.8394
DBSI	0.8075	0.8029	0.8022	0.7578

SLiMs-scoring Methods

***I*-score**

$$\hat{I}(m|X) = -\frac{1}{n} \times \sum_{i=1}^{n} (\frac{1}{l} \times \sum_{j=1}^{l} P(a_{ij}) \times log(P(a_{ij}))) \quad (1)$$

$$log(P(a_{ij})) = \begin{cases} log(1-\varepsilon) & if P(a_{ij}) > 1-\varepsilon \\ log(P(a_{ij})) & otherwise \end{cases} \quad (2)$$

Sliding Window Score (SWS)

$$P(s|X) = \frac{1}{l} \times \sum_{i=1}^{l} P(s_i) \quad (3)$$

We define a threshold λ. If $P(s|X)$ is larger than λ, site s is considered as a real site, and marked as a, otherwise, it is not a site.

$$P(m|X) = \frac{1}{n} \times \sum_{i=1}^{n} P(a_i|X) \quad (4)$$

Position	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	0	0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.6
2	0	0	0	0	0	0	0.3	0	0	0	0	0.6	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0.3	0	0	0	0	0	0.6	0	0	0	0	0	0	0	0	0	0	0	0

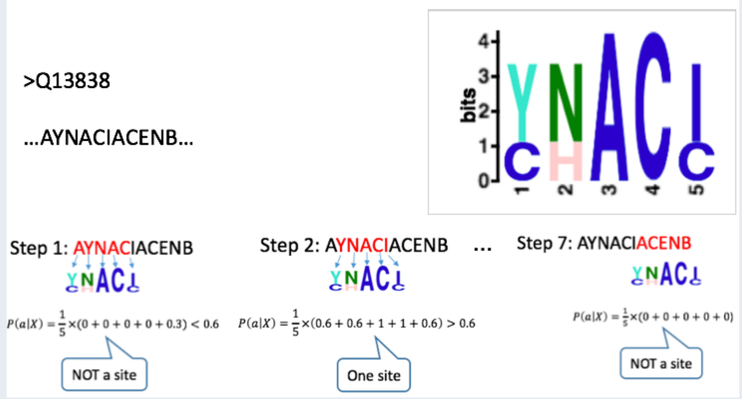

Figure 4: An example of the SWS method.

Conclusion and Future Work

Conclusion

- SLiM features and negative data selection.
- RF performs a little better than SVM.
- PubChem yield better results those of Klekota and Roth.
- Our method outperforms existing methods in terms of AUC performance.

Future work

- Use one-class SVM classification methods and semi-supervised classification methods.
- Combine the fingerprints defined by PubChem and Klekota and Roth together as drug features.

References

[1] Norman E Davey, Niall J Haslam, Denis C Shields, and Richard J Edwards. Slimsearch 2.0: biological context for short linear motifs in proteins. *Nucleic acids research*, 39(suppl 2):W56–W60, 2011.

[2] Yoshihiro Yamanishi, Edouard Pauwels, Hiroto Saigo, and Véronique Stoven. Extracting sets of chemical substructures and protein domains governing drug-target interactions. *Journal of chemical information and modeling*, 51(5):1183–1194, 2011.

[3] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240, 2008.

[4] Dong-Sheng Cao, Shao Liu, Qing-Song Xu, Hong-Mei Lu, Jian-Hua Huang, Qian-Nan Hu, and Yi-Zeng Liang. Large-scale prediction of drug–target interactions using protein sequences and drug topological structures. *Analytica chimica acta*, 752:1–10, 2012.

[5] Zaynab Mousavian, Sahand Khakabimamaghani, Kaveh Kavousi, and Ali Masoudi-Nejad. Drug–target interaction prediction from pssm based evolutionary information. *Journal of pharmacological and toxicological methods*, 78:42–51, 2016.

[6] Yong-Cui Wang, Zhi-Xia Yang, Yong Wang, and Nai-Yang Deng. Computationally probing drug-protein interactions via support vector machine. *Letters in Drug Design & Discovery*, 7(5):370–378, 2010.

[7] Yoshihiro Yamanishi, Masaaki Kotera, Minoru Kanehisa, and Susumu Goto. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, 26(12):i246–i254, 2010.

[8] Mehmet Gönen. Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization. *Bioinformatics*, 28(18):2304–2310, 2012.

[9] Hailin Chen and Zuping Zhang. A semi-supervised method for drug-target interaction prediction with consistency in networks. *PloS one*, 8(5):e62975, 2013.

[10] Feixiong Cheng, Chuang Liu, Jing Jiang, Weiqiang Lu, Weihua Li, Guixia Liu, Weixing Zhou, Jin Huang, and Yun Tang. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol*, 8(5):e1002503, 2012.